機械学習による
データ分析プロセス

鴨志田 亮太 | Ryota Kamoshida

自己紹介

鴨志田 亮太（かもしだ りょうた）

日立製作所　研究開発本部
　旧中央研究所

トップエスイー9期生
　修了制作「MALSS: 機械学習支援ツール」
　データ分析を支援するPythonライブラリを作成
　　https://github.com/canard0328/malss

お話したいこと

正規化, L2ノルム, ニューラルネットワーク, 平均絶対誤差, 名義尺度, 教師あり学習, 量的変数, 決定木, 目的変数, 決定係数, 訓練誤差, 機械学習の分類, 特徴選択, F値, 交差検証, 完全情報最尤推定法, Stratified CV, 学習曲線, 偽陽性率, Feature hashing (Hashing trick), 適合率, 汎化性能, 間隔尺度, 線形モデル, Apriori, Lasso, ロジスティック回帰, バイアス・バリアンス, ノーフリーランチ定理, Under fitting, L1ノルム, ROC曲線, 混合正規分布, リッジ回帰, 次元削減, 機械学習アルゴリズム, ハイパーパラメータ, 次元の呪い, 多重代入法, 分類（識別）, 誤差率, 醜いアヒルの子定理, 精度, 逐次学習, ランダムフォレスト, LOOCV, グリッドサーチ, 順序尺度, 過学習, 混同行列, 再現率, 階層的クラスタリング, 欠損値, 質的変数, k-meansクラスタリング, 回帰, 一般化線形モデル, 比例尺度, 判別分析, 真陽性率, 正則化, 平均二乗誤差, 外れ値検出, 強化学習, SVM, ナイーブベイズ, 主成分分析, k近傍法, 能動学習, AUC, クラスタリング, 頻出パタンマイニング, 教師なし学習, 説明変数（特徴量）, ダミー変数（1-of-K表現）, CRISP-DM

お話すること

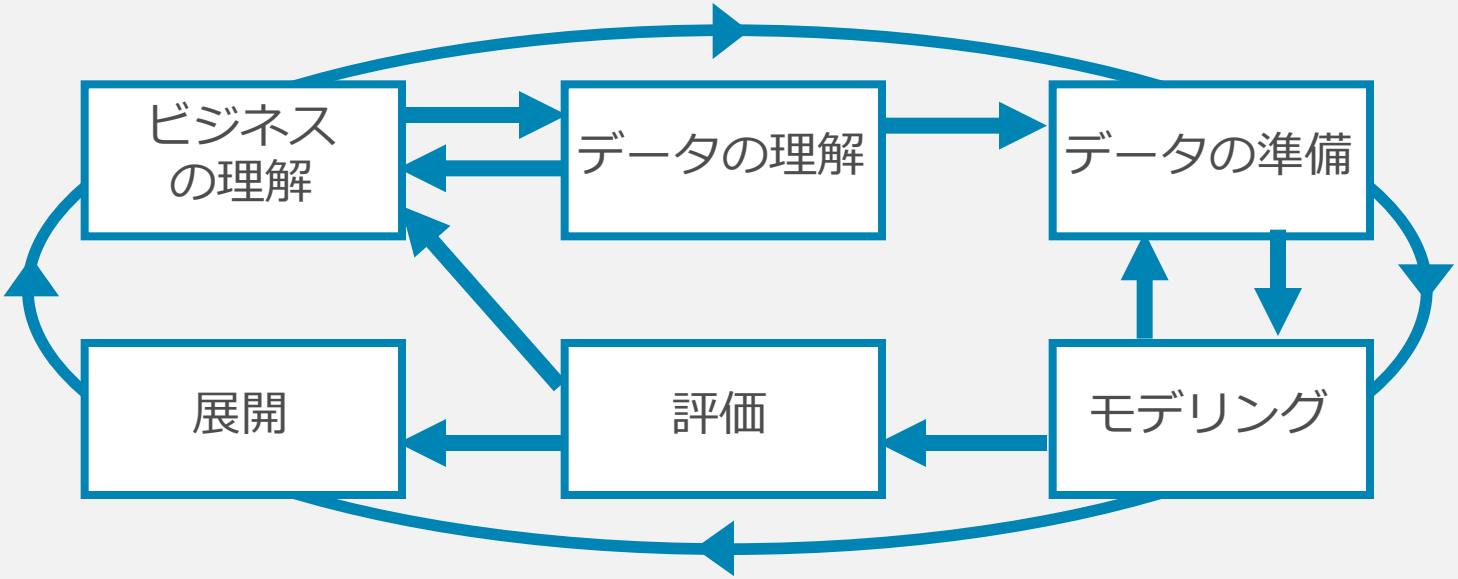
正規化, L2ノルム, ニューラルネットワーク, 平均絶対誤差, 名義尺度, 教師あり学習, 量的変数, 決定木, 目的変数, 決定係数, **訓練誤差**, 機械学習の分類, 特徴選択, F値, **交差検証**, 完全情報最尤推定法, Stratified CV, **学習曲線**, 偽陽性率, Feature hashing (Hashing trick), 適合率, **汎化性能**, 間隔尺度, 線形モデル, Apriori, Lasso, ロジスティック回帰, **バイアス・バリアンス**, **ノーフリーランチ定理**, Under fitting, L1ノルム, ROC曲線, 混合正規分布, リッジ回帰, 次元削減, 機械学習アルゴリズム, ハイパーパラメータ, **次元の呪い**, 多重代入法, 分類（識別）, 誤差率, **醜いアヒルの子定理**, 精度, 逐次学習, ランダムフォレスト, LOOCV, グリッドサーチ, 順序尺度, **過学習**, 混同行列, 再現率, 階層的クラスタリング, 欠損値, 質的変数, k-meansクラスタリング, 回帰, 一般化線形モデル, 比例尺度, 判別分析, 真陽性率, 正則化, 平均二乗誤差, 外れ値検出, 強化学習, SVM, ナイーブベイズ, 主成分分析, k近傍法, 能動学習, AUC, クラスタリング, 頻出パタンマイニング, 教師なし学習, 説明変数（特徴量）, ダミー変数（1-of-K表現）, CRISP-DM

お話しないこと

もう少し詳しい話はこちらに

機械学習によるデータ分析まわりのお話
http://www.slideshare.net/canard0328/ss-44288984

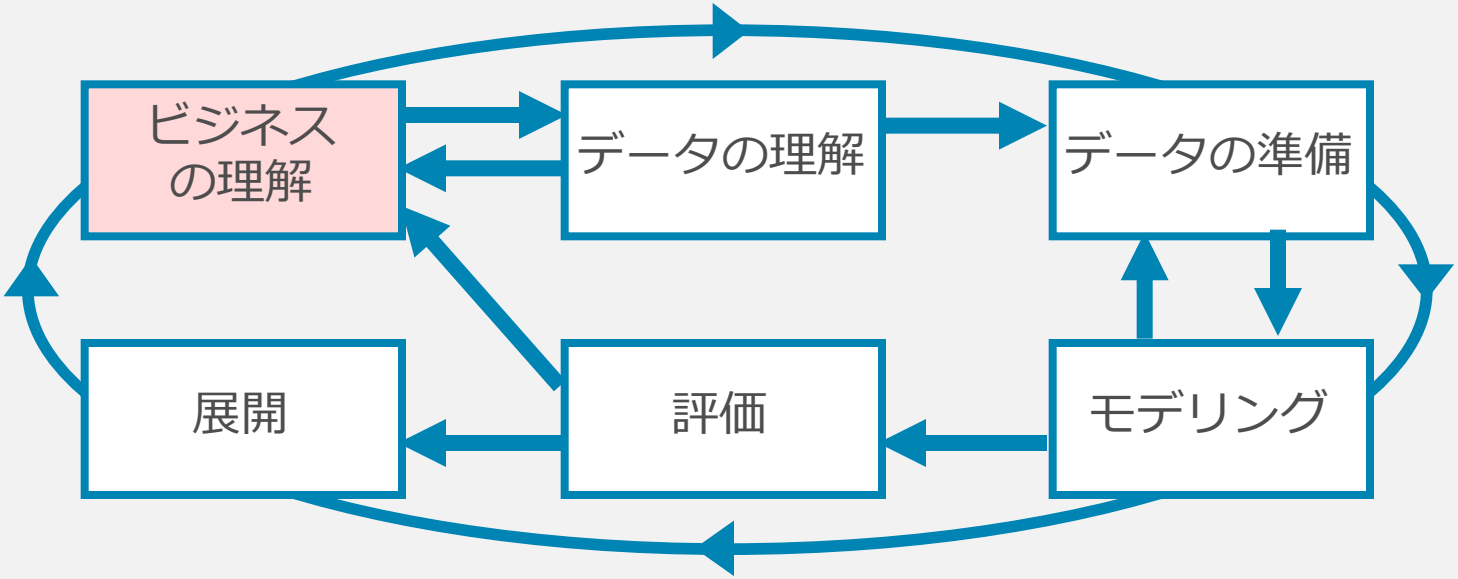
CRISP-DM



CRISP-DM（CRoss-Industry Standard Process for Data Mining）
SPSS，NCR，ダイムラークライスラー，OHRAがメンバーとなっているコンソーシアムで開発されたデータマイニングのための方法論を規定したもの.

（マーケターのためのデータマイニング講座, ITmedia エンタープライズ）

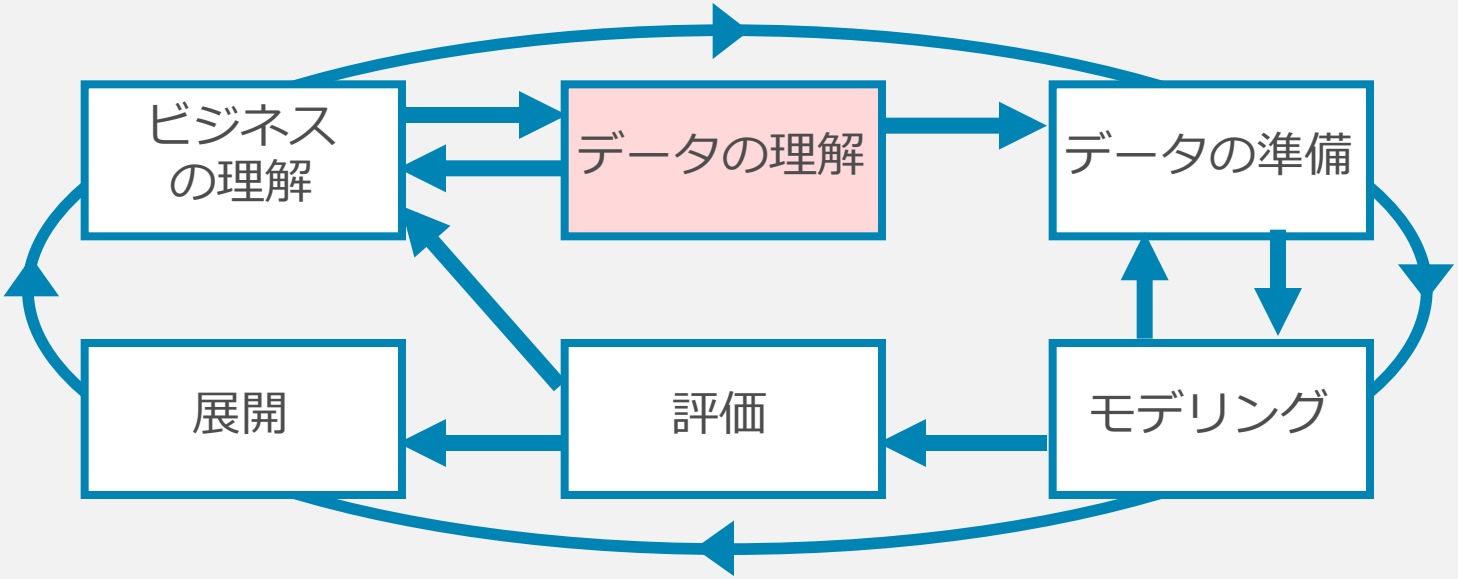
CRISP-DM



ビジネスの理解

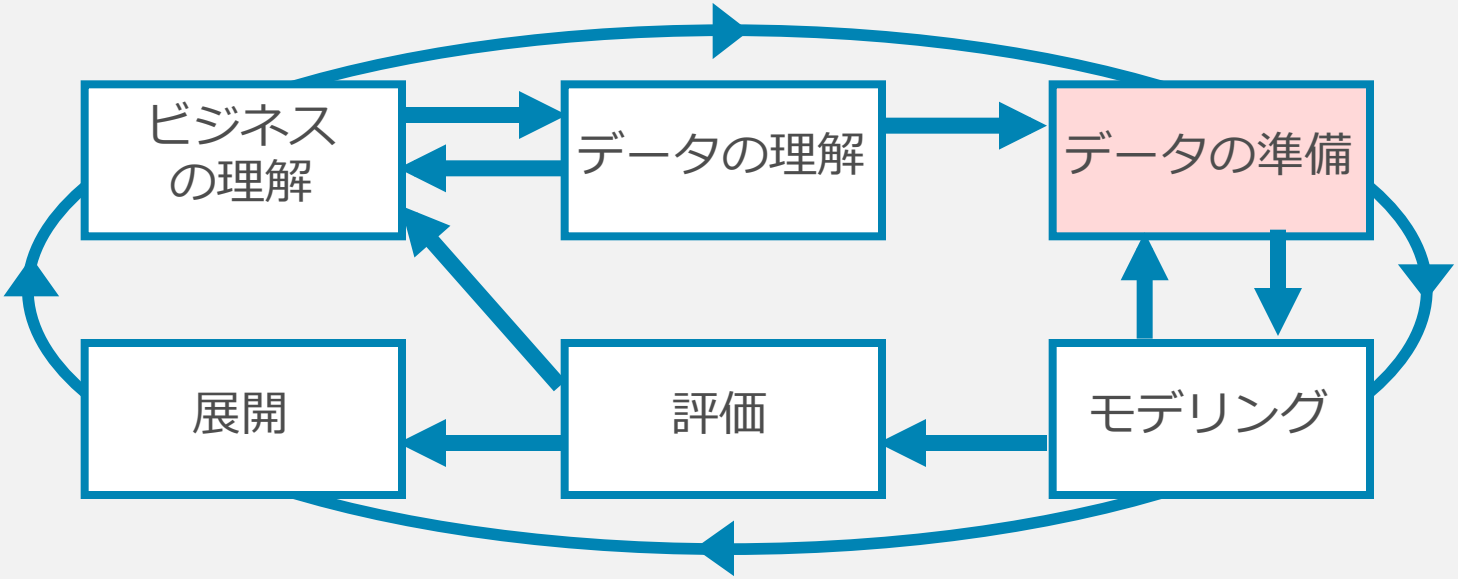
プロジェクト目標の設定を行う.
企業内の各種課題を明確にしたうえで、
データマイニングプロジェクト全体をプランニング
していく.

CRISP-DM



データの理解

どのようなデータが利用可能か，データ項目，量，品質などを調査

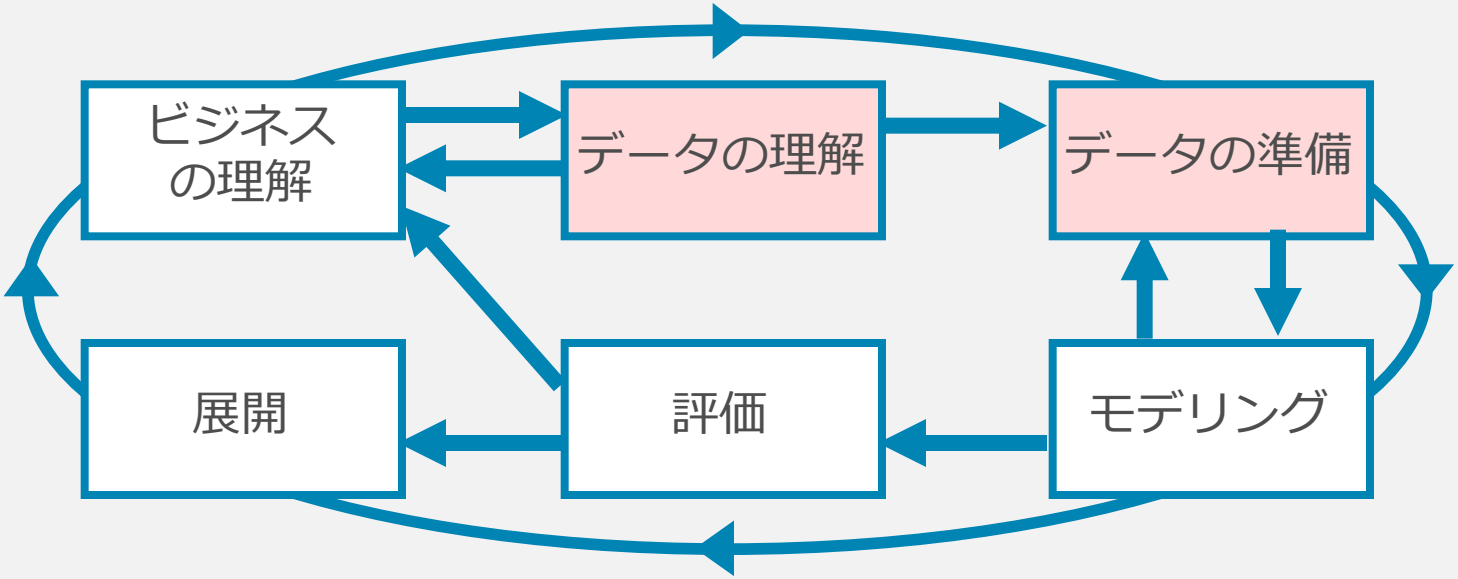


データの準備

使用可能なデータを分析に適したデータに整形（前処理）

欠損値処理，データ型の整理，正規化，サンプリング，etc

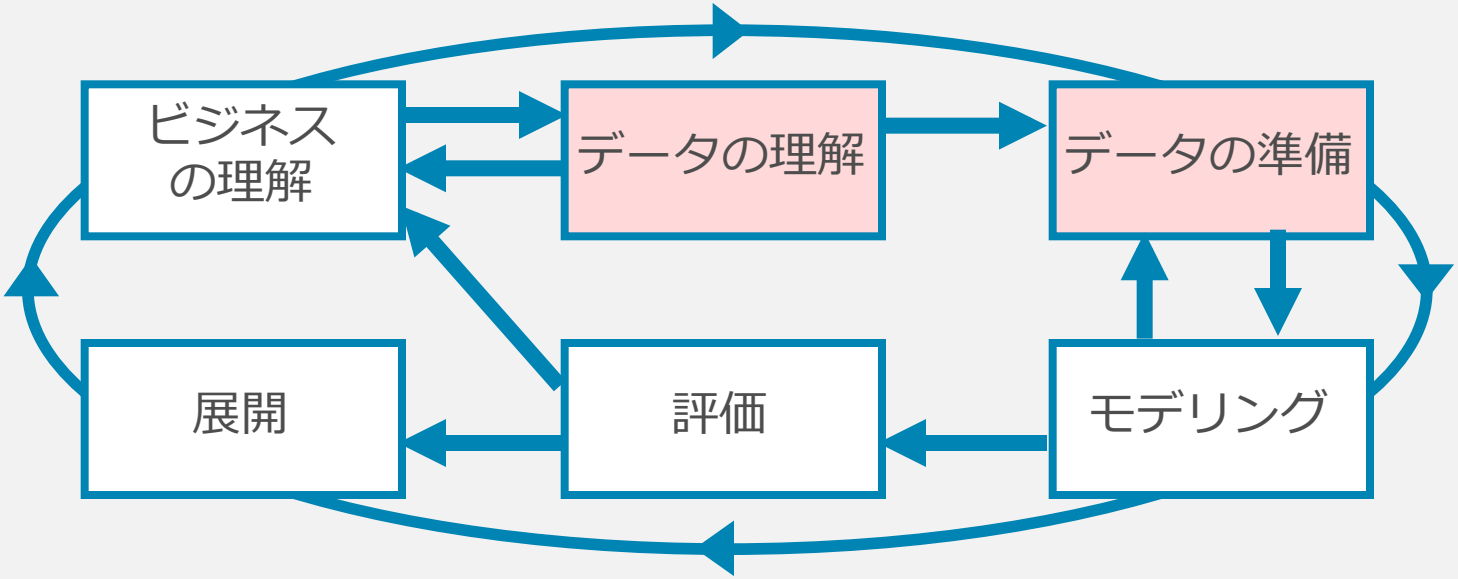
CRISP-DM



データの理解・準備が分析の質を決める

特に特徴量の設計が重要

CRISP-DM



データの理解・準備が分析の質を決める

More than two-thirds of data scientists say cleaning and organizing data is their most-time consuming task and 52.3 percent say that poor quality data is their biggest daily obstacle.

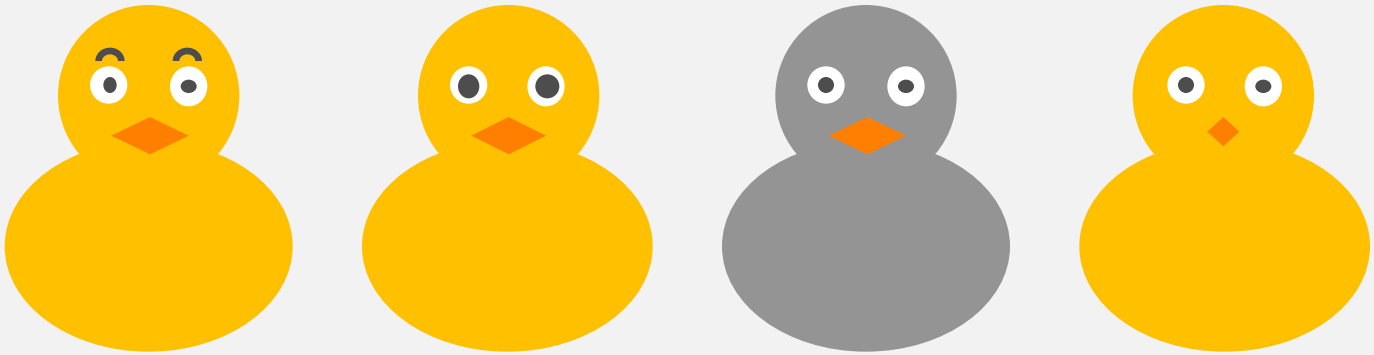
TNW News: Hot to hire data scientists

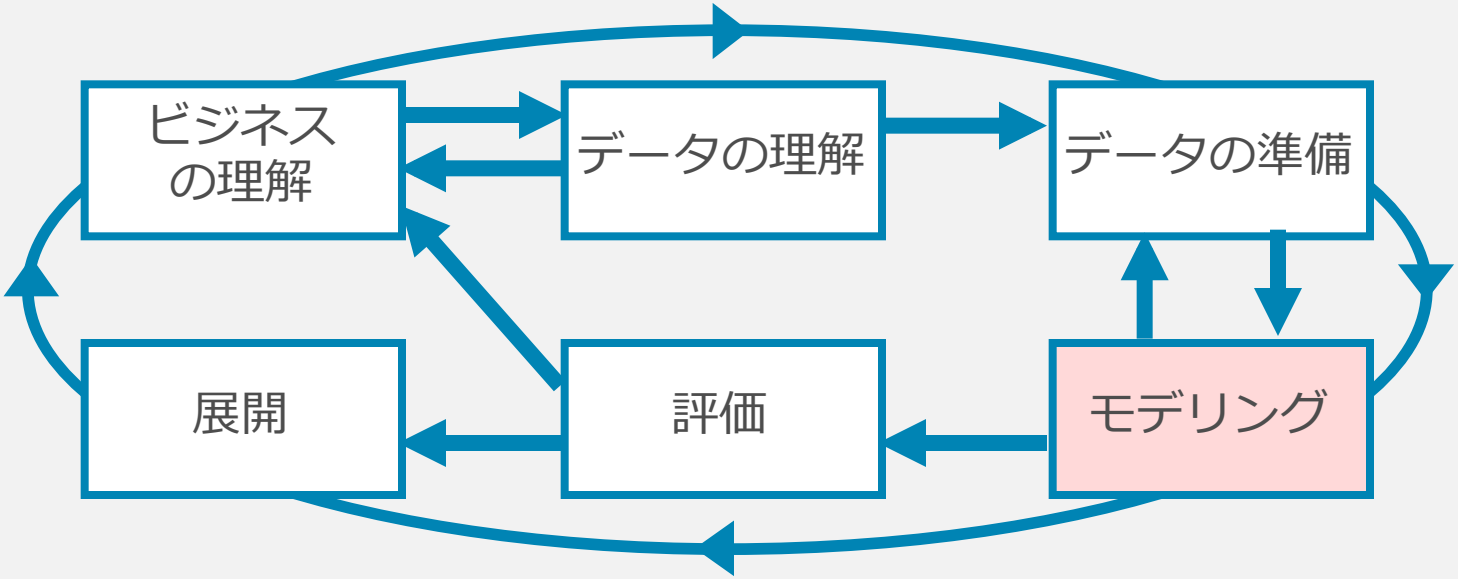
醜いアヒルの子定理(Ugly duckling theorem)

醜いアヒルの子と普通のアヒルの子の類似性は
２羽の普通のアヒルの子の類似性と等しい

問題から独立した**万能な特徴量は存在しない**

特徴量の設計が重要





モデリング

課題を解決するための数理モデルを，仮説に基づいて構築する.

モデル選択→モデリング→評価→前に戻る

アルゴリズム利用時の注意点

ノーフリーランチ定理

あらゆる問題で性能の良い

万能な学習アルゴリズムは存在しない

目的に適したアルゴリズムを選択しましょう

とは言っても，実用上上手くいくことの多い，少数のアルゴリズムが頻繁に利用されるのも事実

次元の呪い

次元の呪い(Curse of dimensionality)
特徴量（説明変数）の数が増えると汎化性能※を向上させることが難しくなる

使えそうなデータはなんでも特徴量に加えてしまえ，は危険

特徴選択や**次元削減**により特徴量の数を減らす

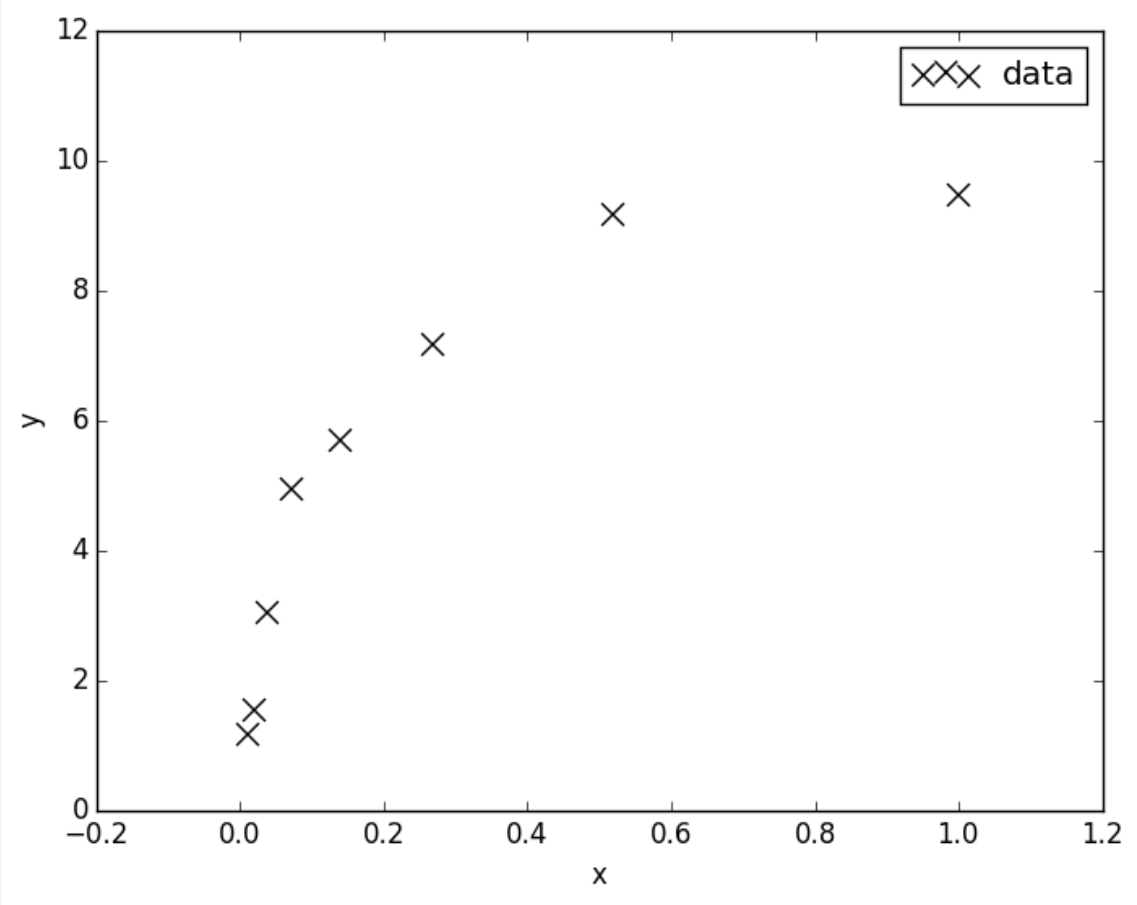
データを用意する段階で特徴量を吟味することが非常に重要

次元の呪いについて，詳しくは「球面集中現象」を検索

※未知のデータを予測する性能

過学習

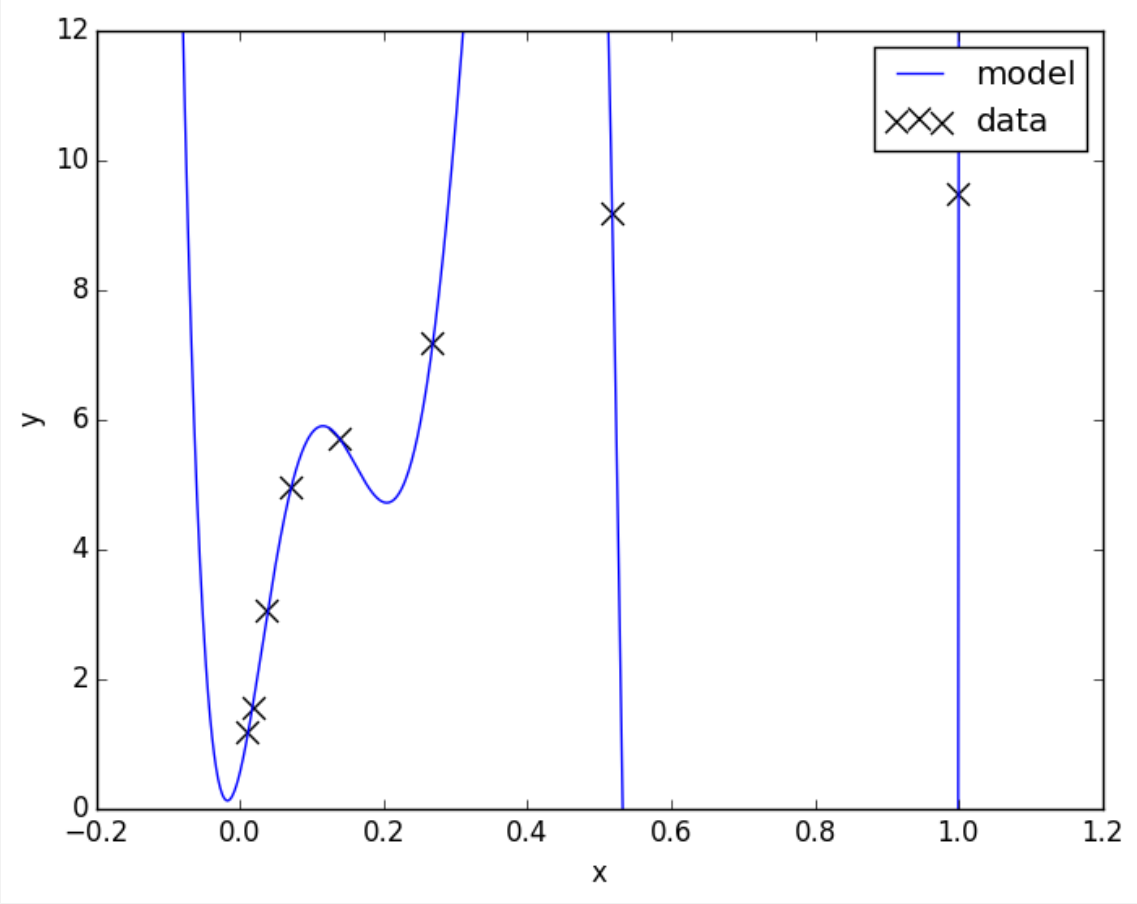
xの値からyの値を予測するモデルを作りたい



過学習

出来た！誤差０！完璧！！

・・・本当ですか！？



過学習

過学習(Over fitting)

与えられたデータに（ノイズも含めて）過度に適合してしまい，**訓練誤差**は小さいが，未知データに対する性能が低下してしまう状態.

汎化性能

未知のデータに対する性能（汎化性能）を定量化した**汎化誤差**を小さくすることが重要

表現力の高いアルゴリズム使用時，特徴量が多いとき，与えられたデータが少ないときに過学習しやすい.

過学習(Over fitting)

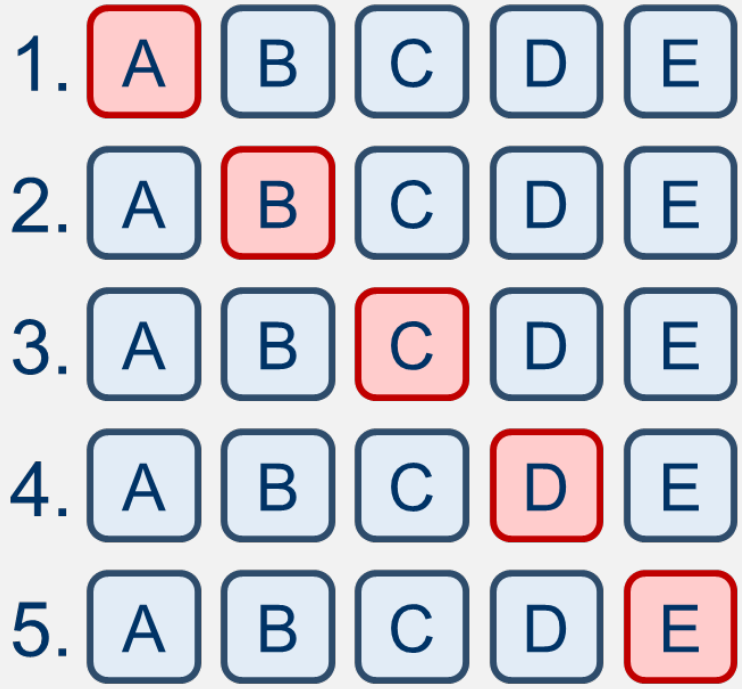
与えられたデータに（ノイズも含めて）過度に適合してしまい，**訓練誤差**は小さいが，未知データに対する性能が低下してしまう状態.

百度は認められている以上のテストを行う不正を働いたと報告した。イメージネットは、テスト参加社に対し、1週間に2回テストを受けることを認めているが、同社が同日ブログに掲載した説明では、百度は3月に5日間で40回以上のテストを受けるなど、6カ月間のテスト回数が約200回に達した。

THE WALL STREET JOURNAL 人工知能テスト結果で謝罪―中国・百度

交差検証(Cross validation)

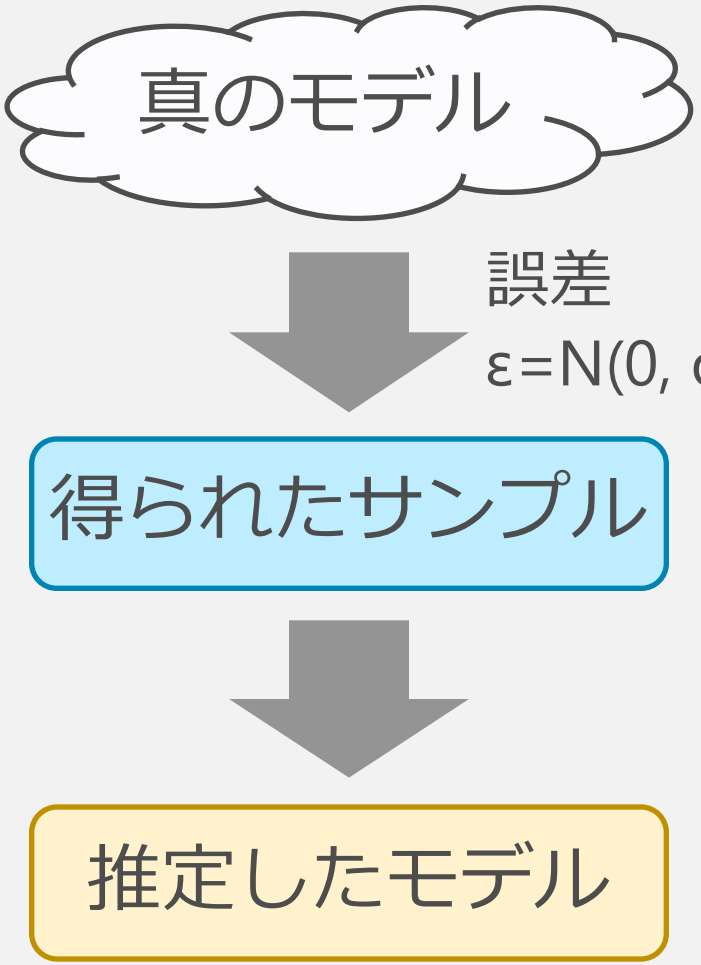
データを学習用と評価用に分割する



1. B～Eで学習，Aで評価
2. A,C～Eで学習，Bで評価
3. A,B,D,Eで学習，Cで評価
4. A～C,Eで学習，Dで評価
5. A～Dで学習，Eで評価
6. １～5の平均を算出

5分割交差検証（5-fold cross validation）

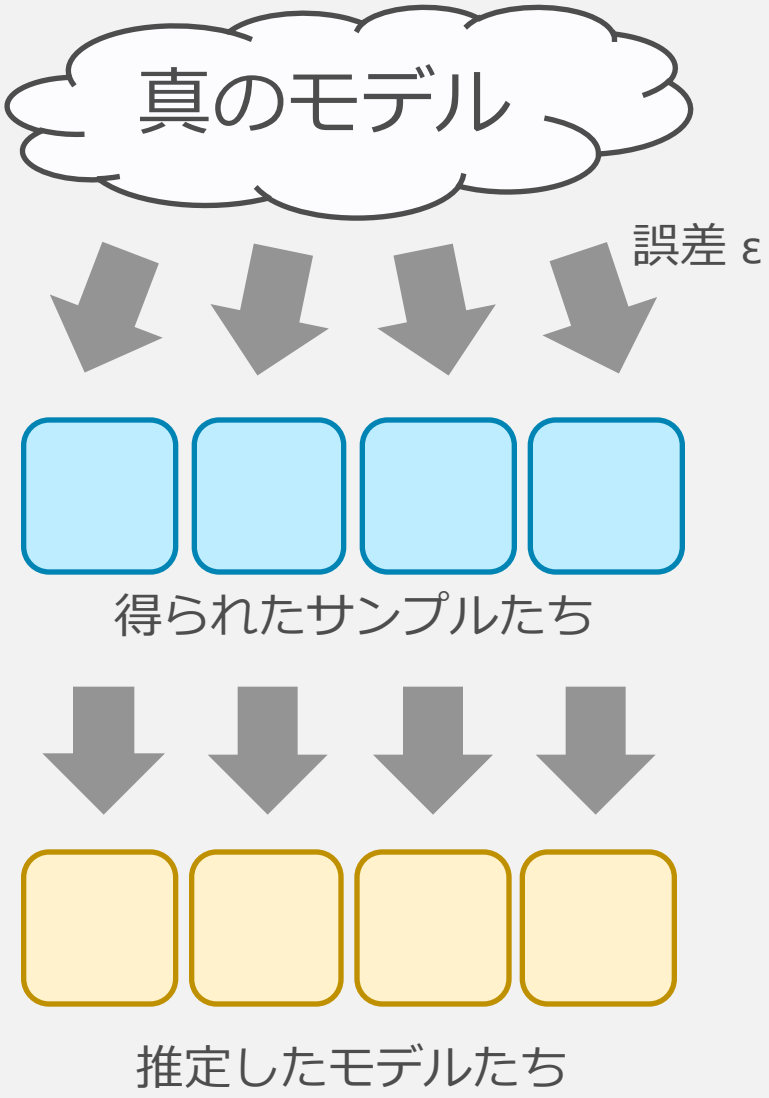
誤差について

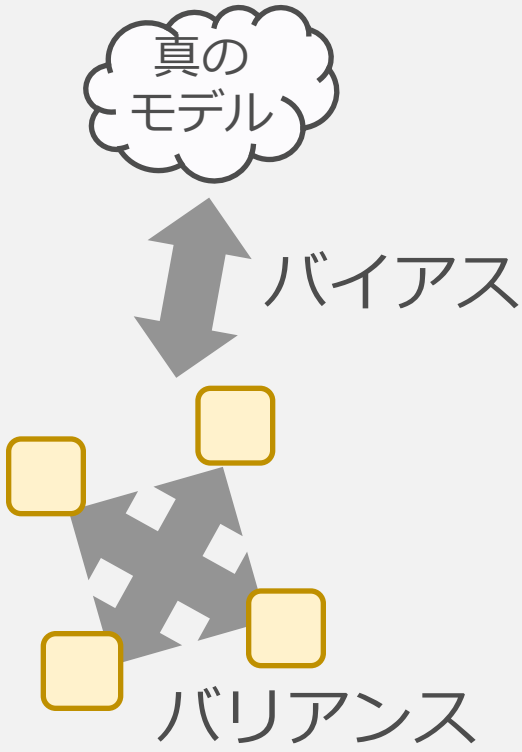


$$差 = \sigma^2 + Bias^2 + Variance$$

Bias(バイアス)
　モデルの能力に起因する差
Variance(バリアンス)
　サンプルに起因する差

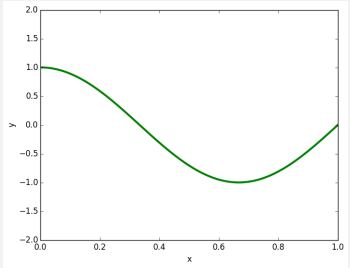
バイアスとバリアンス





バイアスとバリアンス

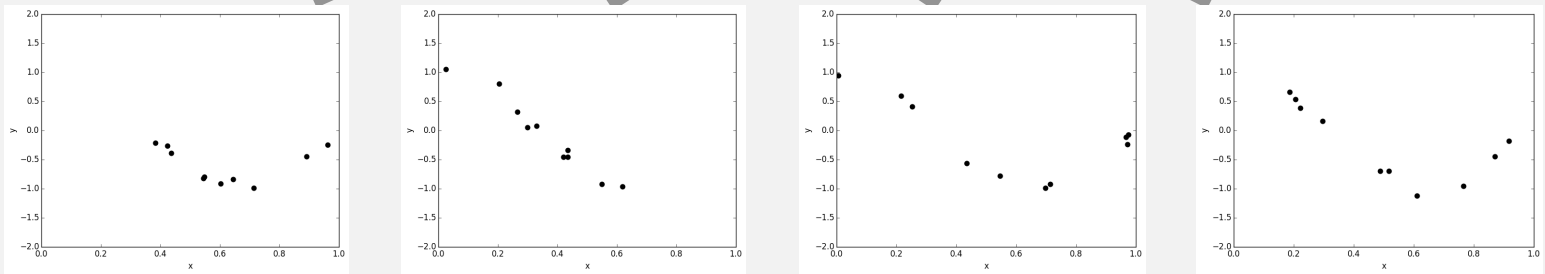
1次式でモデリング



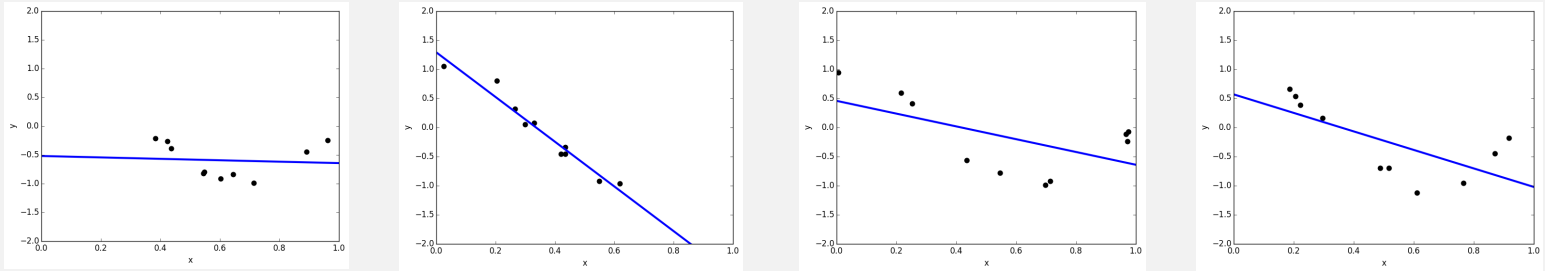
真のモデル

誤差 ε

得られた
サンプルたち



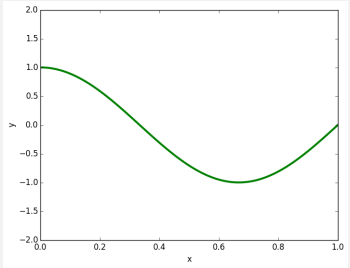
推定した
モデルたち



差は大きいが，差のばらつきは小さい　→　**ハイバイアス／ローバリアンス**

バイアスとバリアンス

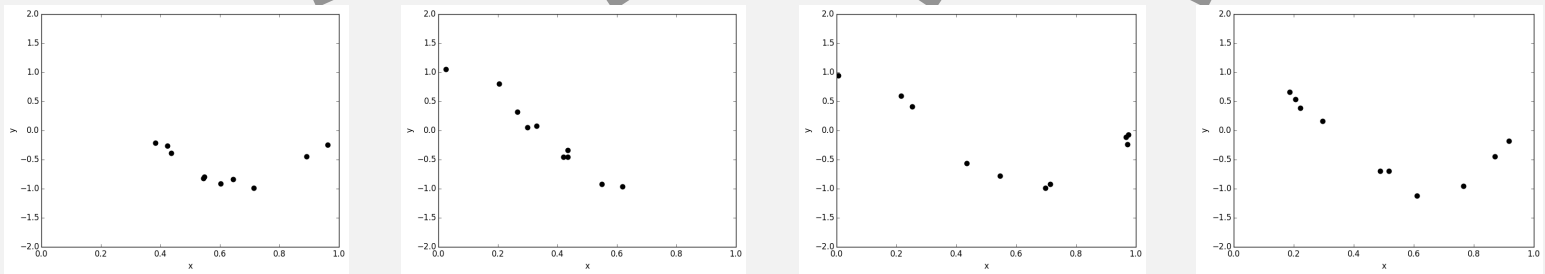
多項式でモデリング



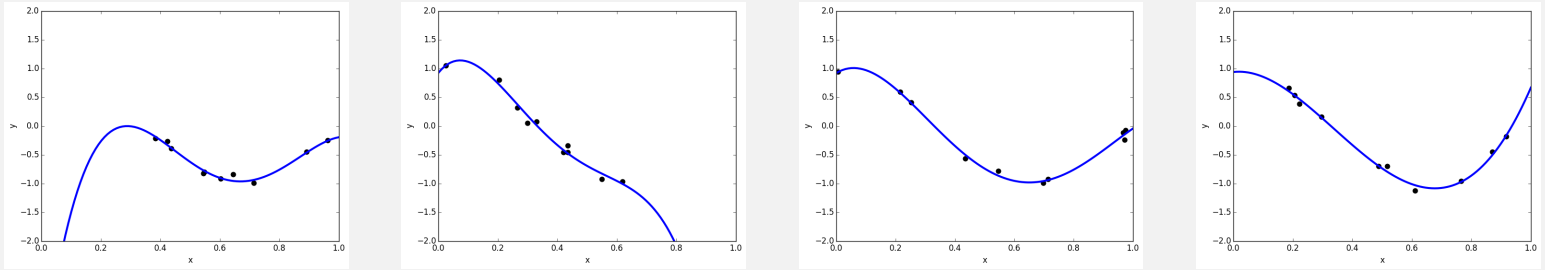
真のモデル

誤差 ε

得られた
サンプルたち



推定した
モデルたち



サンプルによる差が大きい　→　**ローバイアス／ハイバリアンス**

バイアスとバリアンス

バイアスとバリアンスは**トレードオフ**の関係

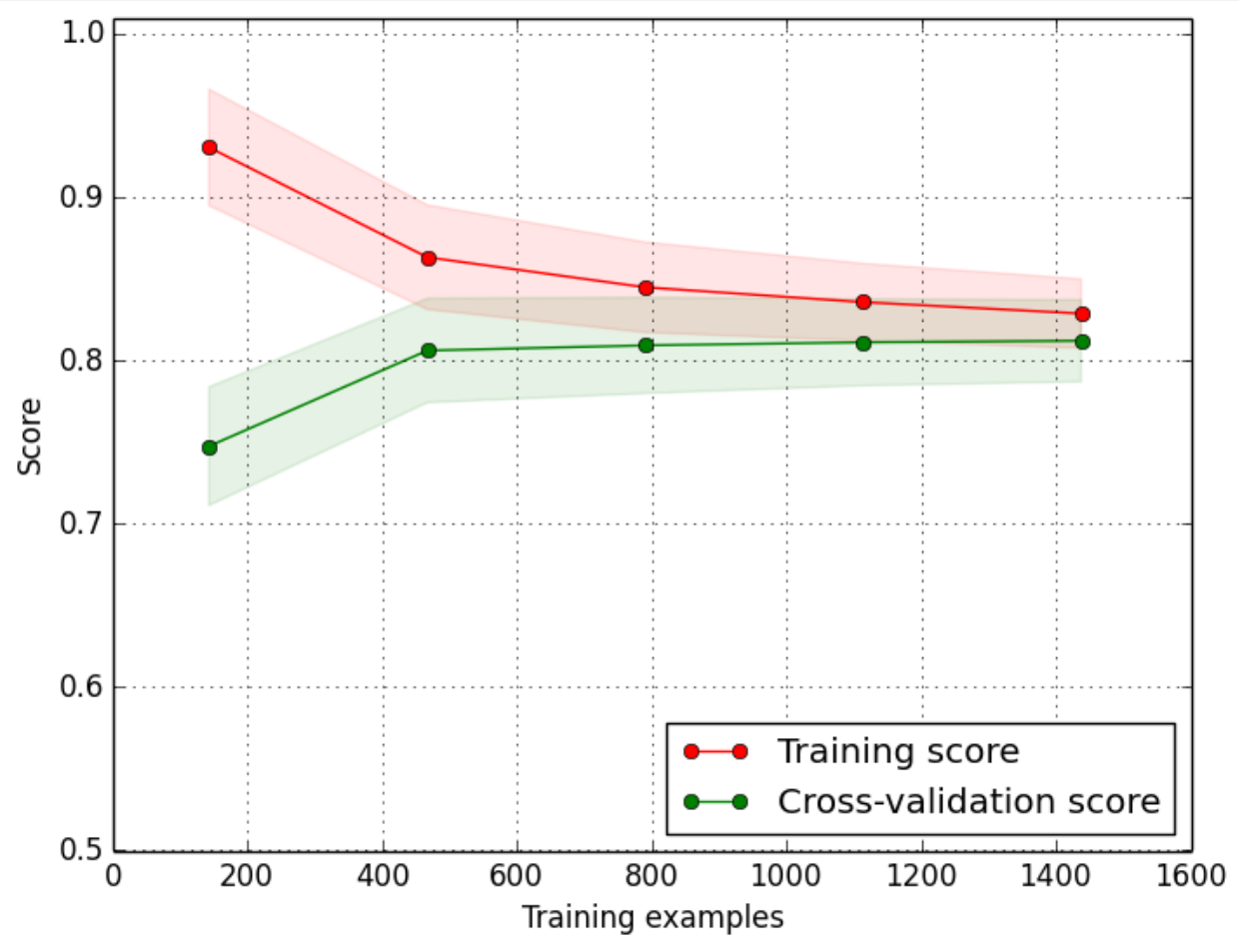
柔軟性の高いモデル（アルゴリズム）
バイアス小，バリアンス大⇒**ハイバリアンス**
過学習(Over fitting)

柔軟性の低いモデル（アルゴリズム）
バイアス大，バリアンス小⇒**ハイバイアス**
Under fitting

現在のモデルの状態を確認するには？

学習曲線

データサイズを変えながら訓練スコア(誤差)
汎化スコア(誤差)をプロット



ハイバイアスの目安

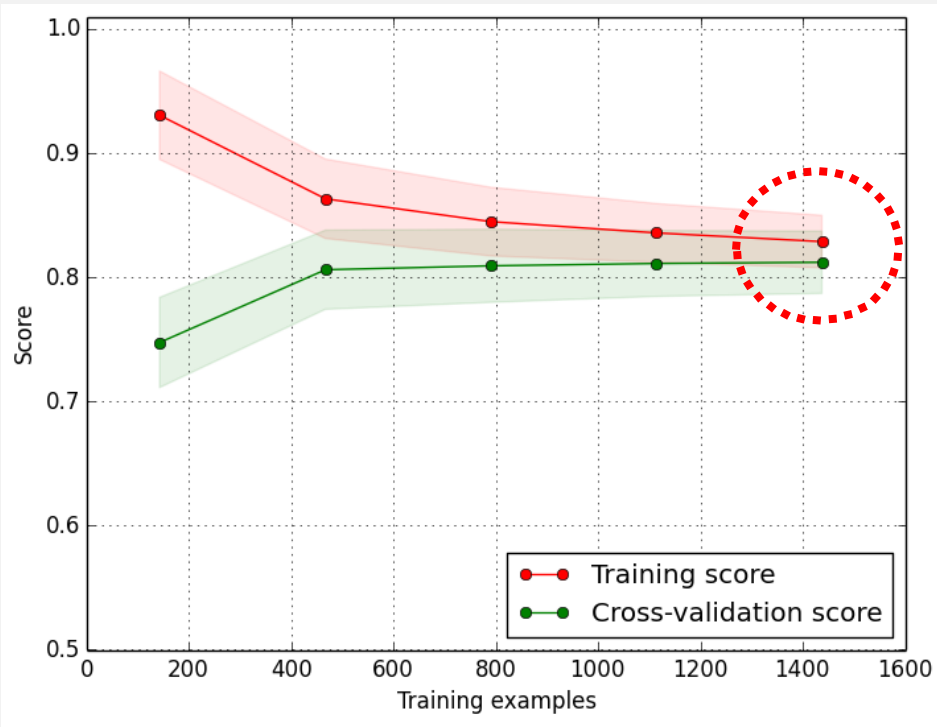
訓練スコア(誤差)が低い(大きい)
訓練スコアと汎化スコアの差が小さい

ハイバリアンスの目安

訓練スコアと汎化スコアの差が大きい
汎化スコアの改善がサチっていない

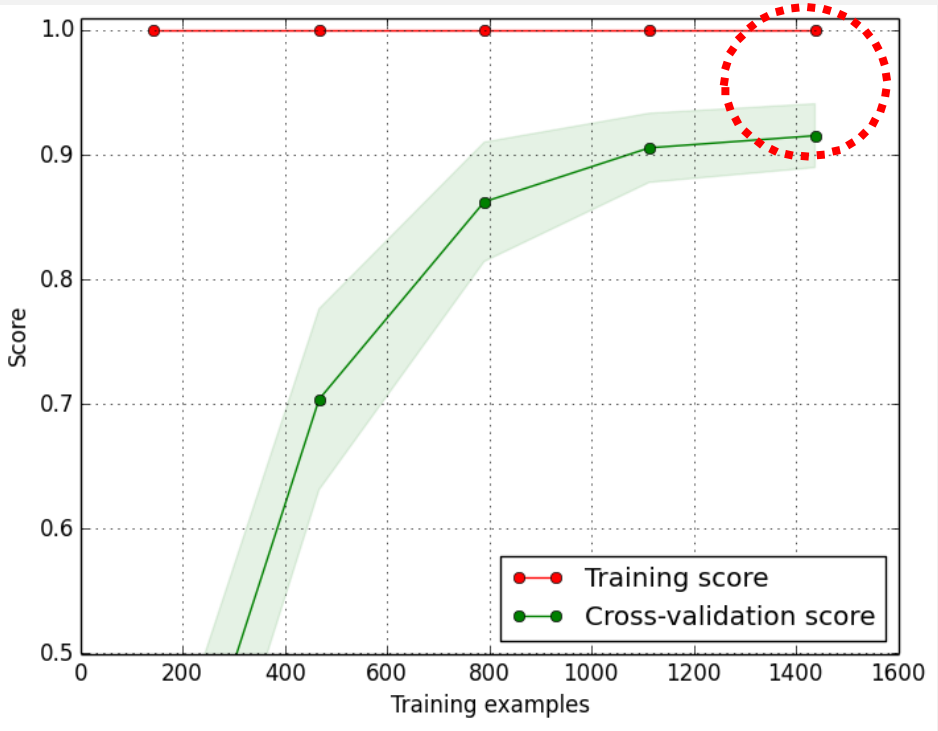
学習曲線

ハイバイアス



スコアが低い
スコアの差が小さい

ハイバリアンス



スコアの差が大きい

ハイバイアスの場合

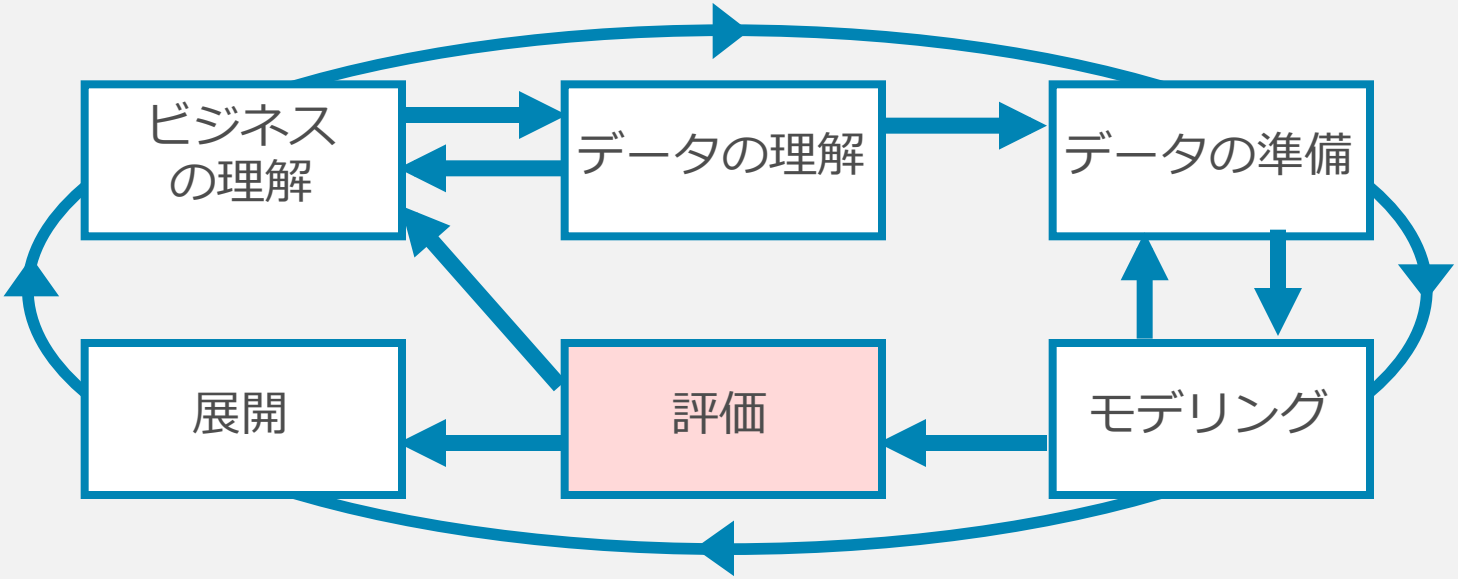
（有効な）特徴量を増やす

アルゴリズムを（柔軟性の高いものに）変更する

ハイバリアンスの場合

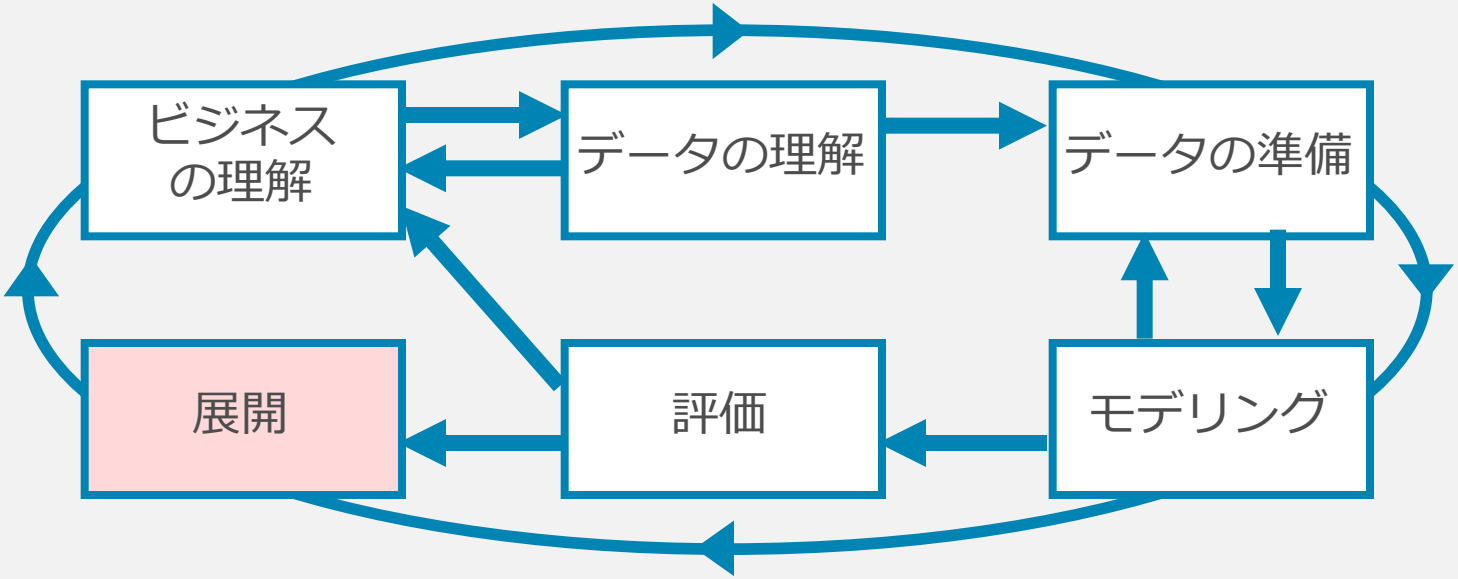
データを増やす

（不要な）特徴量を削除する



評価

「ビジネスの理解」で定義したビジネス目標を達成するに十分なモデルであるかを**ビジネスの観点から**評価する.



展開

データ分析した結果をビジネスに適用するための
具体的なプランニングを行っていく.

本日お話したこと

1. データの理解・準備が分析の質を決める
2. 醜いアヒルの子定理
3. ノーフリーランチ定理
4. 次元の呪い
5. 過学習
6. バイアス・バリアンス